Chapter 11: Models for Matched Pairs

Dipankar Bandyopadhyay

Department of Biostatistics,
Virginia Commonwealth University

BIOS 625: Categorical Data & GLM

[Acknowledgements to Tim Hanson and Haitao Chu]

Example (Table 11.1 p. 414): Presidential Votes in 2004 and in 2008, for males sampled by the General Social Survey. The 433 males are cross classified according to their two (binary) responses (X, Y):

	2008 Election		
2004 Election	Democrat	Republican	Total
Democrat	175	16	191
Republican	54	188	242
Total	229	204	433

- Here, each person is matched with himself. This is also called *repeated measures* data.
- Here we see people tend to select Democrat/Republican both times more often than change their opinion.
- Question: of those that change their opinion, which direction do they tend to go? Hint: $\frac{16}{16+175} \approx 0.08$ & $\frac{54}{54+188} \approx 0.22$

Let $\pi_{ab} = P(X = a, Y = b)$ and n_{ab} be the number of such pairs.

	2008 Election	
2004 Election	Democrat $Y = 1$	Republican $Y = 2$
Democrat $X = 1$	π_{11} & n_{11}	π_{12} & n_{12}
Republican $X = 2$	π_{21} & n_{21}	π_{22} & n_{22}

- We assume $(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{mult}\{n_{++}, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})\}$.
- When $\pi_{1+} = \pi_{+1}$ then $P(X = 1) = P(Y = 1)$ and we have *marginal homogeneity*. This is of course equivalent to $P(X = 2) = P(Y = 2)$ by looking at complimentary events.
- In the Presidential Votes data, this would indicate that the proportion of people that select Democrat at 2004 is equal to the proportion that select Democrat at 2008. Does it imply that no one has changed their mind?

Let $p_{ab} = n_{ab}/n_{++}$ be the sample proportion in each cell. Define the difference $\delta = \pi_{+1} - \pi_{1+} = P(Y = 1) - P(X = 1)$.
What does this measure for the Presidential Votes data? δ is estimated by

$$d = p_{+1} - p_{1+} = \frac{n_{11} + n_{21} - (n_{11} + n_{12})}{n_{++}}.$$

Considering the covariance for multinomial vector elements, we have a $(1 - \alpha)100\%$ CI for δ is

$$d \pm z_{\alpha/2}\hat{\sigma}(d),$$

where

$$\hat{\sigma}(d) = \sqrt{[(p_{12} + p_{21}) - (p_{12} - p_{21})^2]/n_{++}}.$$

To test $H_0 : \delta = 0$, i.e. $H_0 : P(X = 1) = P(Y = 1)$, the Wald test statistic is $z_0 = d/\hat{\sigma}(d)$. The score test statistic is

$$z_0 = \frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}}.$$

- A p-value for testing $H_0 : \delta = 0$ is $P(|Z| > |z_0|)$; this latter test is *McNemar's test*.
- For the Presidential Votes data, $\hat{\delta} = (54 - 16)/433 = 0.088$ with a 95% CI for δ is $(0.051, 0.125)$. The number of people selecting Democrat has increased by by 8.8% with a 95% CI of $(5.1\%, 12.5\%)$. The McNemar (score) test statistic for testing $H_0 : P(X = 1) = P(Y = 1)$ is $z_0 = 4.542 (= \sqrt{20.63})$ yielding a p-value of ≤ 0.001.
- Does this mean that between 5.1% and 12.5% of the people have changed their minds? (Answer: no).

- By having a person serve as his own control we increase the precision with which this difference is estimated (relative to two *iid* samples at 2004 and 2008).
- In some sense it is easier to measure how peoples attitudes are changing by looking directly at changes within an individual instead of considering separate populations at 2004 and 2008.
- Note that

$$n \text{ var}(d) = \pi_{1+}(1-\pi_{1+}) + \pi_{+1}(1-\pi_{+1}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}).$$

When the response is positively correlated, $\pi_{11}\pi_{22} > \pi_{12}\pi_{21}$ and the variance is smaller relative to two independent samples.

- McNemar's test statistic is *not* a function of diagonal elements, but the sample difference d and $\hat{\sigma}(d)$ are.
- The diagonal elements contribute to how correlated Y_{i1} and Y_{i2} are, i.e. the tendency for people to not change their mind: $P(Y_{i1} = Y_{i2} = 1) = n_{11}/n_{++}$ and $P(Y_{i1} = Y_{i2} = 2) = n_{22}/n_{++}$.
- Of those that *make a switch*, the off-diagonal elements determine the direction and strength of the switch.
- We may be interested in how the odds of selecting Democrat change for a randomly selected individual from the population (conditional inference)
- or, we may be interested in how the odds of selecting Democrat change across the the two populations: everyone at 2004, and everyone at 2008 (marginal inference).

- We can recast this as a *marginal* logit model

$$\text{logit } P(Y_{ij} = 1) = \mu + \boldsymbol{\beta}' \mathbf{x}_{ij},$$

  where $\mathbf{x}_{i1} = 0$ and $\mathbf{x}_{i2} = 1$ are "before" and "after" covariates. For the PV example, the covariates represent time (2004 vs. 2008).
- In general, $\mathbf{x}_{ij}$ are any covariates of interest, but the correlation between Y_{i1} and Y_{i2}, $\alpha = \text{corr}(Y_{i1}, Y_{i2})$ must be accounted for in some way in estimating $\boldsymbol{\beta}$. For the PV example, this correlation is quite high, the polychoric (tetrachoric) correlation is estimated to be $\hat{\rho} = 0.904$ with $\hat{\sigma}(\hat{\rho}) = 0.023$.
- We will discuss marginal categorical models that account for such correlation, or *clustering*, fit via GEE in Chapter 12.
- When fitting this type of model in GENMOD, $\hat{\beta} = log[(229/204/(191/242)] = 0.352$ and so $e^{\hat{\beta}} = 1.42$. $\widehat{\text{corr}}(Y_{i1}, Y_{i2}) = \hat{\alpha} = 0.69$.

Let (Y_{i1}, Y_{i2}) be a pair of ordered responses from the i^{th} subject, $i = 1, \ldots, n$. Consider

$$\text{logit } P(Y_{ij} = 1) = \alpha_i + \beta x_j,$$

where $x_1 = 0$ and $x_2 = 1$. Here, $j = 1, 2$ can be thought of as time, with Y_{i1} denoting the first observation taken on subject i and Y_{i2} being the second. Then

$$\frac{P(Y_{i1} = 1)}{P(Y_{i1} = 0)} = e^{\alpha_i} \text{ and } \frac{P(Y_{i2} = 1)}{P(Y_{i2} = 0)} = e^{\alpha_i} e^{\beta}.$$

And so

$$\theta_{21} = \frac{P(Y_{i2} = 1)/P(Y_{i2} = 0)}{P(Y_{i1} = 1)/P(Y_{i1} = 0)} = e^{\beta},$$

which does not depend on the subject i.

- The $\alpha_1, \ldots, \alpha_n$ are subject-specific effects that correlate Y_{i1} and Y_{i2}. Large α_i indicates that *both* $Y_{i1} = 1$ and $Y_{i2} = 1$ are likely. Small α_i indicates that *both* $Y_{i1} = 0$ and $Y_{i2} = 0$ are likely.
- The *model* assumes that given the $\alpha_1, \ldots, \alpha_n$, the responses are independent. That is, $Y_{i1} \perp Y_{i2}$, independent across all $i = 1, \ldots, n$.
- An estimate of e^{β} provides a conditional odds ratio. For a given person, the odds of success are e^{β} more likely at time $j = 2$ over time $j = 1$. It is conditional on the value of α_i, i.e. the person.
- When $\alpha_1 = \alpha_2 = \cdots = \alpha_n$ then there is no person-to-person variability in the response pair (Y_{i1}, Y_{i2}). The pairs (Y_{i1}, Y_{i2}) are then *iid* from the population.

The joint mass function for the n pairs $\{(Y_{11}, Y_{12}), \ldots, (Y_{n1}, Y_{n2})\}$ is given by

$$\prod_{i=1}^{n} \left(\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \right)^{y_{i1}} \left(\frac{1}{1+e^{\alpha_i}} \right)^{1-y_{i1}} \left(\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \right)^{y_{i2}} \left(\frac{1}{1+e^{\alpha_i+\beta}} \right)^{1-y_{i2}}.$$

The pairwise success totals $S_i = y_{i1} + y_{i2} \in \{0, 1, 2\}$ are sufficient for α_i. We can compute (see book Section 11.2.3)

$$\begin{aligned}
P(Y_{i1} = 0, Y_{i2} = 0 | S_i = 0) &= 1 \\
P(Y_{i1} = 1, Y_{i2} = 1 | S_i = 2) &= 1 \\
P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) &= \frac{e^{\beta}}{1+e^{\beta}} \\
P(Y_{i1} = 1, Y_{i2} = 0 | S_i = 1) &= \frac{1}{1+e^{\beta}}
\end{aligned}$$

Conditional inference is based on conditioning on $\{S_1, \ldots, S_n\}$. Let $n_{12} = \sum_{i=1}^{n} I\{Y_{i1} = 1, Y_{i2} = 0\}$, $n_{21} = \sum_{i=1}^{n} I\{Y_{i1} = 0, Y_{i2} = 1\}$, and $n^* = n_{12} + n_{21}$ are the total number with $S_i = 1$. The conditional likelihood is

$$\prod_{i:S_i=1} \left(\frac{e^{\beta}}{1+e^{\beta}}\right)^{y_{i1}} \left(\frac{1}{1+e^{\beta}}\right)^{y_{i2}} = \frac{[e^{\beta}]^{n_{21}}}{[1+e^{\beta}]^{n^*}}.$$

It pleasantly turns out that $\hat{\beta} = \log(n_{21}/n_{12})$ and $\hat{\sigma}(\hat{\beta}) = \sqrt{1/n_{21} + 1/n_{12}}$.

- PV data: We have $\hat{\beta} = \log(54/16) = 1.22$ and $\hat{\sigma}(\hat{\beta}) = 0.285$. So the odds of a randomly selected person selecting Democrat at 2008 is estimated to be $e^{1.22} = 3.38$ times their initial odds at 2004.
- An alternative approach to conditioning on sufficient statistics is to specify a full model and treat the α_i as subject-specific random effects. If we can think of subjects as being exchangeable, then a common assumption is

$$\alpha_1, \ldots, \alpha_n \overset{iid}{\sim} N(\mu, \sigma^2).$$

There are only three parameters (μ, σ, β) in the likelihood (after averaging out the $\alpha_1, \ldots, \alpha_n$). Studies have shown that estimating β is robust to the distributional assumption placed on $\alpha_1, \ldots, \alpha_n$. More to come in Chapter 13.

Generalize to repeated measures within a cluster

We can think of taking two or more observations within a cluster (an individual, matched covariates, etc.)

Let (Y_{i1}, Y_{i2}) be a pair of correlated binary observations from within the same cluster. The data look like

Y_{i1}	Y_{i2}	$\mathbf{x}_{i1}$	$\mathbf{x}_{i2}$
Y_{11}	Y_{12}	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
Y_{21}	Y_{22}	$\mathbf{x}_{21}$	$\mathbf{x}_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{n1}	Y_{n2}	$\mathbf{x}_{n1}$	$\mathbf{x}_{n2}$

The logit model specifies

$$\text{logit } P(Y_{ij} = 1) = \alpha_i + \mathbf{x}'_{ij}\boldsymbol{\beta},$$

where $i = 1, \ldots, n$ is a *pair* number and $j = 1, 2$ denotes the observation within a cluster.

As before, we condition on the sufficient statistics for β, namely $S_i = Y_{i1} + Y_{i2}$. We have

$$\begin{aligned} P(Y_{i1} = Y_{i2} = 0 | S_i = 0) &= 1 \\ P(Y_{i1} = Y_{i2} = 1 | S_i = 2) &= 1 \\ P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) &= \exp(\mathbf{x}'_{i2}\beta)/[\exp(\mathbf{x}'_{i1}\beta) + \exp(\mathbf{x}'_{i2}\beta)] \\ P(Y_{i1} = 1, Y_{i2} = 0 | S_i = 1) &= \exp(\mathbf{x}'_{i1}\beta)/[\exp(\mathbf{x}'_{i1}\beta) + \exp(\mathbf{x}'_{i2}\beta)]. \end{aligned}$$

The conditional likelihood is formed as before in the simpler case and inference obtained in PROC LOGISTIC using the STRATA statement. Let's examine the PV data using thinking of (Y_{i1}, Y_{i2}) as repeated measurements within an individual with corresponding covariates $x_{i1} = 0$ and $x_{i2} = 1$ denoting time.

```
data Data1;
 do ID=1 to 175;
    dem=1; time=0; output;
    dem=1; time=1; output; end;
 do ID=176 to 191;
    dem=1; time=0; output;
    dem=0; time=1; output; end;
 do ID=192 to 245;
    dem=0; time=0; output;
    dem=1; time=1; output; end;
 do ID=246 to 433;
    dem=0; time=0; output;
    dem=0; time=1; output; end;
proc logistic data=Data1;
 strata ID;
 model dem(event='1')=time;
 run;
```

The LOGISTIC Procedure

Conditional Analysis

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.7844	1	<.0001
Score	20.6286	1	<.0001
Wald	18.2627	1	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
time	1	1.2164	0.2846	18.2627	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
time	3.375	1.932	5.896

Matched case-control studies

Let $(Y_{i1} = 0, Y_{i2} = 1)$ be a pair of binary observations from two different subjects matched on criteria that could affect the outcome. The data look like

Control Y_{i1}	Case Y_{i2}	Case $\mathbf{x}_{i1}$	Control $\mathbf{x}_{i2}$
0	1	$\mathbf{x}_{11}$	$\mathbf{x}_{12}$
0	1	$\mathbf{x}_{21}$	$\mathbf{x}_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	1	$\mathbf{x}_{n1}$	$\mathbf{x}_{n2}$

The logit model specifies

$$\text{logit } P(Y_{ij} = 1) = \alpha_i + \mathbf{x}'_{ij}\boldsymbol{\beta},$$

where $i = 1, \ldots, n$ is a *pair* number and $j = 1, 2$ denotes case or control.

By construction we have all $S_i = y_{i1} + y_{i2} = 1$ and analogous to our conditional approach for a pair of binary responses within an individual, we have

$$P(Y_{i1} = 0, Y_{i2} = 1 | S_i = 1) = \frac{e^{\mathbf{x}'_{i2}\boldsymbol{\beta}}}{e^{\mathbf{x}'_{i1}\boldsymbol{\beta}} + e^{\mathbf{x}'_{i2}\boldsymbol{\beta}}},$$

which does not depend on α_i, and the conditional likelihood for $\boldsymbol{\beta}$ is formed by taking the product over $i = 1, \ldots, n$.

Even though the number of cases and the number of controls are fixed at n, the logit link allows us to determine the effect of covariates on the *odds* of being a case versus a control. That is the odds of being a case instead of a control is increased by e^{β_j} when x_j is increased by unity.

Example (p. 422): $n_{++} = 144$ pairs of Navajo Indians, one having myocardial infarction (MI) and the other free of heart disease, were matched on age and gender yielding 288 Navajo total. It is of interest to determine how the presence of diabetes affects the odds of MI. Here's the cross-classification of the *pairs*:

	MI cases	
MI controls	Diabetes	No diabetes
Diabetes	9	16
No diabetes	37	82

The data are conditionally analyzed using the STRATA subcommand in PROC LOGISTIC.

```
data Data1;
 do ID=1 to 9; case=1; diab=1; output; case=0; diab=1; output; end;
 do ID=10 to 25; case=1; diab=0; output; case=0; diab=1; output; end;
 do ID=26 to 62; case=1; diab=1; output; case=0; diab=0; output; end;
 do ID=63 to 144; case=1; diab=0; output; case=0; diab=0; output; end;
proc logistic data=Data1;
  strata ID;
  model case(event='1')=diab;
  run;
```

The LOGISTIC Procedure

Conditional Analysis

Model Information

Response Variable	case
Number of Response Levels	2
Number of Strata	144
Model	binary logit

Probability modeled is case=1.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
diab	1	0.8383	0.2992	7.8501	0.0051

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
diab	2.312	1.286	4.157

We estimate that the odds of MI increase by 2.3 when diabetes is present, with a 95% CI of $(1.3, 4.2)$. Diabetes significantly affects the outcome MI.

The following data is from Breslow and Day (1980) and is analyzed in the SAS documentation. There's 63 matched pairs, consisting of one case of endometrial cancer (Outcome=1) and a control without cancer (Outcome=0). The case and corresponding control have the same ID, specified in the `strata` subcommand. Two prognostic factors are included: Gall (= 1 for gall bladder disease) and Hyper (= 1 for hypertension). The goal of the case-control analysis is to determine the relative risk of endometrial cancer for gall bladder disease, controlling for the effect of hypertension.

```
data d1;
 do ID=1 to 63; do Outcome = 1 to 0 by −1; input Gall Hyper @@; output; end; end;
  datalines ;
0 0  0 0  0 0  0 0  0 1  0 1  0 0  1 0  1 0  0 1  0 1  0 0  1 0  0 0  1 1  0 1  0 0  0 0  0 0  0 0
1 0  0 0  0 0  0 1  1 0  0 1  1 0  1 0  1 0  0 1  0 1  0 0  0 0  1 1  0 0  1 1  0 0  0 1  0 1  0 0
0 0  1 1  0 1  0 1  0 1  0 0  0 0  0 0  0 0  0 0  0 0  0 1  1 0  0 1  0 0  0 1  1 0  0 0  0 1  0 0
0 1  0 0  0 1  0 0  0 1  0 0  0 0  0 0  1 1  1 1  0 0  0 1  0 1  0 0  0 1  0 1  0 1  0 1  0 1  0 0
0 0  0 0  0 1  1 0  0 0  0 1  0 0  0 0  1 0  0 0  0 0  0 0  1 1  0 0  0 1  0 0  0 0  0 0  0 1  0 1
0 0  0 0  0 1  0 1  0 1  0 0  0 1  0 0  1 0  0 0  0 0  0 0  1 1  1 0  0 0  0 0  0 0  0 0  1 1  0 0
1 0  1 0  0 1  0 0  1 0  0 0
;
proc  logistic  data=d1; strata  ID;
 model outcome(event='1')= Gall Hyper; run;
```

The LOGISTIC Procedure

Conditional Analysis

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.5487	2	0.1029
Score	4.3620	2	0.1129
Wald	4.0060	2	0.1349

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Gall	1	0.9704	0.5307	3.3432	0.0675
Hyper	1	0.3481	0.3770	0.8526	0.3558

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Gall	2.639	0.933	7.468
Hyper	1.416	0.677	2.965

Adjusting for hypertension, the odds of developing endometrial cancer are about 2.6 times as great (and almost significant!) for those with gall bladder disease. How about the relative risk?

- Generalization: more than a pair of binary outcomes, $j = 1, 2, \ldots, J_i$. For example, repeated measures on subject i, or J_i rats from litter i.
- Section 11.1 presented marginal inference. Answers how does probability marginally change $\delta = P(Y = 1) - P(X = 1)$, averaged over everyone in population.
- Section 11.2 deals with a conditional interpretation. θ_{21} was how odds of success change over time $j = 2$ versus $j = 1$ for any randomly sampled *individual* in the population.
- In matched case-control study, we use the α_i to induce correlation in responses (Y_{i1}, Y_{i2}) within two like individuals.
- For sparse data, one can include an additional EXACT subcommand in PROC LOGISTIC to get exact tests and odds ratio estimates, e.g.

```
exact diab / estimate=both;
```

Final comment on PV data:

- The conditional odds ratio 3.38 is bigger (further away from the null) than the population averaged odds ratio 1.42. Is this reasonable? Yes. Many people either select Democrat or Republican. If one's $\alpha_i << 0$ then this person select Republican regardless of β. After 4 years, this person perhaps likes the Democrat a bit more, but the probability in either case is likely to be small.
- Which inference is preferred? It depends on the question!
  - The conditional inference holds for an individual with repeated measures, or individuals in a matched (*blocked*!) set. Because the conditional approach essentially blocks on like variables (measurements within an individual; outcomes matched on gender, age, cholesterol, etc.) it accounts for, and can reduce variability associated with estimating the effect of interest.
  - The marginal inference holds for the population as a whole, averaged over the blocking effects.

11.4 Testing for symmetry in a square $I \times I$ table

Consider an $I \times I$ table which cross-classifies (X, Y) on the same outcomes.

	$Y=1$	$Y=2$	$\cdots$	$Y=I$
$X=1$	π_{11}	π_{12}	$\cdots$	π_{1I}
$X=2$	π_{21}	π_{22}	$\cdots$	π_{2I}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$X=I$	π_{I1}	π_{I2}	$\cdots$	π_{II}

Marginal homogeneity happens when $P(X=i) = P(Y=i)$ $(\pi_{+i} = \pi_{+i})$ for $i = 1, \ldots, I$. This is important, for example, when determining if classifiers (like X-ray readers) tend to classify in roughly the same proportions. If not, perhaps one reader tends to diagnose a disease more often than another reader.

Symmetry, a stronger assumption, implies marginal homogeneity.

Definition of symmetric table

An $I \times I$ table is *symmetric* if $P(X = i, Y = j) = P(X = j, Y = i)$ $(\pi_{ij} = \pi_{ji})$.

This simply reduces the number of parameters from I^2 (subject to summing to one) to $I(I+1)/2$ (subject to summing to one). For example, in a 3×3 table this forces

	$Y=1$	$Y=2$	$Y=3$
$X=1$	π_1	π_2	π_3
$X=2$	π_2	π_4	π_5
$X=3$	π_3	π_5	π_6

subject to $\pi_1 + \pi_4 + \pi_6 + 2\pi_2 + 2\pi_3 + 2\pi_5 = 1$.

The symmetric model is easily fit by specifying the cell probabilities by hand in GENMOD. A test of the symmetric model versus the saturated model is a test of $H_0 : \pi_{ij} = \pi_{ji}$ and can be carried out by looking at the Deviance statistic (yielding a LRT).

Recent example

The following table is from Yule (1900)

	Wife		
Husband	Tall	Medium	Short
Tall	18	28	14
Medium	20	51	28
Short	12	25	9

Let (X, Y) be the heights of the (Husband, Wife). The table is symmetric if $P(X = i, Y = j) = P(X = j, Y = i)$. For example, symmetry forces the same proportion of pairings of (Husband,Wife)=(Tall,Short) and (Husband,Wife)=(Short,Tall). This assumes the following structure

	Wife		
Husband	Tall	Medium	Short
Tall	π_1	π_2	π_3
Medium	π_2	π_4	π_5
Short	π_3	π_5	π_6

subject to $\pi_1 + 2\pi_2 + 2\pi_3 + \pi_4 + 2\pi_5 + \pi_6 = 1$.

SAS code

```
data hw;
 input h w symm count @@;
 datalines;
 1 1 1 18   1 2 2 28   1 3 3 14
 2 1 2 20   2 2 4 51   2 3 5 28
 3 1 3 12   3 2 5 25   3 3 6  9
;
proc genmod; class symm;
 model count=symm / link=log dist=poi;
```

The GENMOD output gives us

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3	1.6635	0.5545

A test of symmetry versus the saturated model gives a p-value of $P(\chi^2_3 > 1.66) = 0.65$. We accept that the symmetric model fits.

Symmetry implies marginal homogeneity, $P(X = i) = P(Y = i)$. Husbands and wives are tall, medium, or short in the same proportions.

Furthermore, for example, short wives and tall husbands occur with the same probability as tall wives with short husbands.

11.3 & 11.5 $I \times I$ marginal homogeneity & kappa statistic

Consider an $I \times I$ table where X and Y are cross-classified on the same scale. Below are $n = 118$ slides classified for carcinoma of the uterine cervix by two pathologists as (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma *in situ*, or (4) squamous or invasive carcinoma.

	Pathologist B				
Pathologist A	1	2	3	4	Total
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4	0	1	17	10	28
Total	27	12	69	10	118

If A and B were the *same person* then $\pi_{ij} = 0$ when $i \neq j$, i.e. there'd only be nonzero diagonal elements. Nonzero off-diagonal elements reflect disagreement and the further off the diagonal they are, the more severe the disagreement.

Marginal homogeneity

For example there are two slides classified by B as carcinoma *in situ* (not metastasized beyond the original site) that A classified as negative.

Perfect agreement occurs when $\pi_{11} + \pi_{22} + \pi_{33} + \pi_{44} = 1$. The strength of agreement has to do with how close this is to one.

Marginal homogeneity occurs when the two classifiers agree on the proportion of each classification in the population, but not necessarily the classifications themselves. If marginal homogeneity is not satisfied, then one classifier tends to classify a fixed category more often than the other.

Kappa statistic

Classifiers are independent if $P(X = i, Y = j) = P(X = i)P(Y = j)$, and in this case agreement for category i happens with probability $P(X = i, Y = i) = P(X = i)P(Y = i) = \pi_{i+}\pi_{+i}$. The kappa statistic looks at the difference between the probability of agreement $\sum_{i=1}^{I} \pi_{ii}$ and agreement due to "chance" $\sum_{i=1}^{I} \pi_{i+}\pi_{+i}$, normalized by the largest this can be when $\sum_{i=1}^{I} \pi_{ii} = 1$:

$$\kappa = \frac{\sum_{i=1}^{I} \pi_{ii} - \pi_{i+}\pi_{+i}}{1 - \sum_{i=1}^{I} \pi_{i+}\pi_{+i}},$$

and is estimated by simply replacing π_{ij} by $\hat{\pi}_{ij} = n_{ij}/n_{++}$.

SAS code & output

```
data table;
 input A B count @@;
 datalines;
1 1  22  1 2   2  1 3   2  1 4   0
2 1   5  2 2   7  2 3  14  2 4   0
3 1   0  3 2   2  3 3  36  3 4   0
4 1   0  4 2   1  4 3  17  4 4  10
;
proc freq  order=data; weight count; tables A*B / plcorr agree;
```

The FREQ Procedure

Statistic	Value	ASE
Gamma	0.9332	0.0340
Polychoric Correlation	0.9029	0.0307

Test of Symmetry

Statistic (S)	30.2857
DF	6
Pr > S	<.0001

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.4930	0.0567	0.3818	0.6042
Weighted Kappa	0.6488	0.0477	0.5554	0.7422

Sample Size = 118

Interpretation

- There's a test for symmetry! The statistic is the same as the Pearson GOF test for the symmetric log-linear model, i.e. a score test for testing $H_0 : \pi_{ij} = \pi_{ji}$. What do we conclude?
- How about $\hat{\gamma} = 0.93$ and $\hat{\rho} = 0.90$, both highly significant? What does that tell us?
- Finally, $\hat{\kappa} = 0.49$ with 95% CI about $(0.4, 0.6)$. The difference between observed agreement and that expected purely by chance is between 0.4 and 0.6, moderately strong agreement.
- The weighted kappa statistic is valid for an ordinal response and weights differences in classifications according to how "severe" the discrepancy. See p. 435.
- κ is *one number* summarizing agreement. It may be much more interesting to quantify *where* or *why* disagreement occurs via models.

Test of marginal homogeneity

Recall that McNemar's test tests $H_0 : P(X = 1) = P(Y = 1)$ for a 2×2 table. This is output from `PROC FREQ` in `SAS` using AGREE.

Often, when comparing raters, we have more than 2 categories. A general test of marginal homogeneity tests $H_0 : P(X = i) = P(Y = i)$ for $i = 1, \ldots, I$. `mh` is a small program written by John Uebersax to perform overall tests of marginal homogeneity, among other things.

```
MH Program:  Marginal Homogeneity Tests for N x N Tables
Version 1.2 - John Uebersax
2008-04-24   2:19 PM
 ***INPUT***
  Diagnoses of Carcinoma (Agresi Table 10.8)
  4 categories
  Path A  is row variable
  Path B  is column variable
  ordered categories
      22     2     2     0
       5     7    14     0
       0     2    36     0
       0     1    17    10
  Total number of cases:      118
```

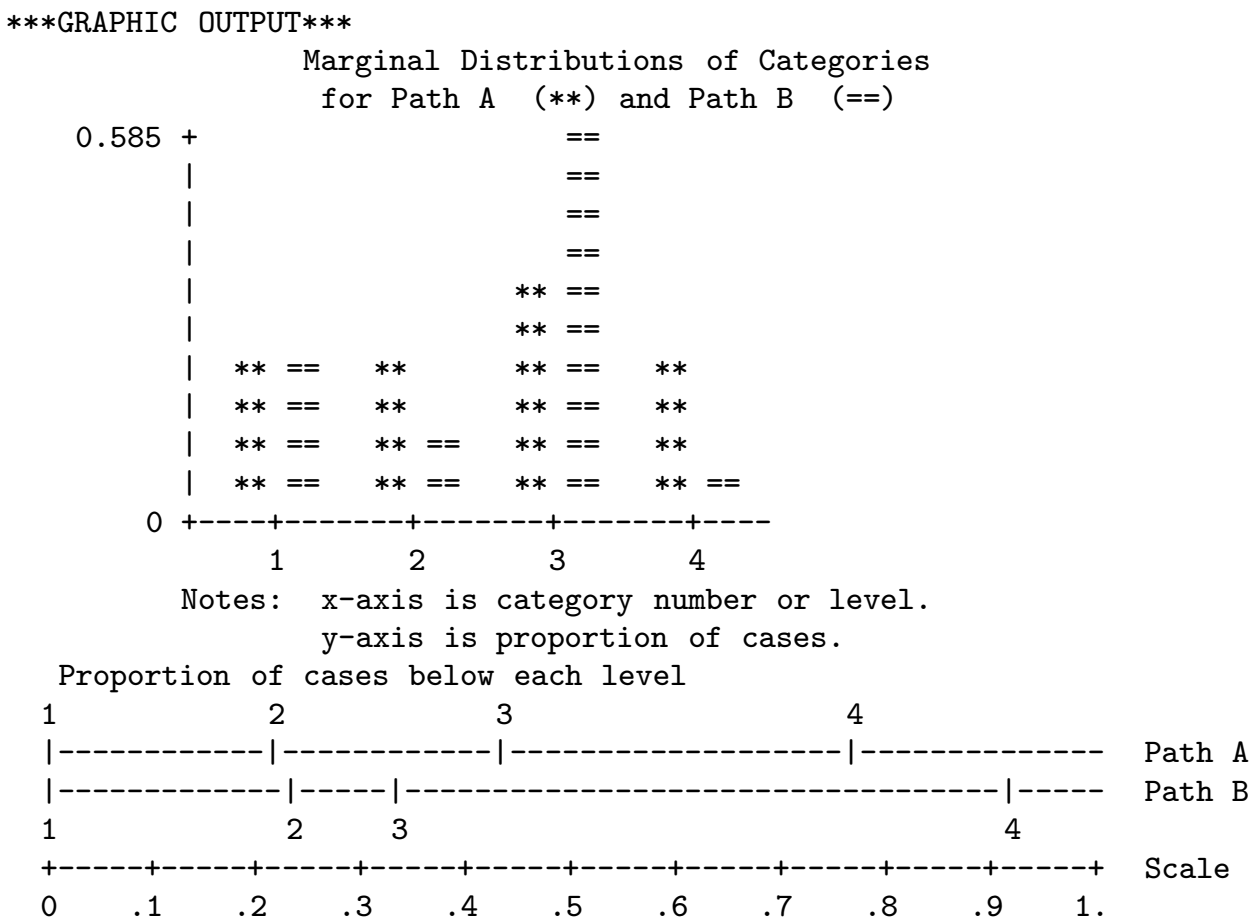
Output

```
***BASIC TESTS***
  Four-fold tables tested
     22      4      5     87
      7     19      5     87
     36      2     33     47
     10     18      0     90
                    McNemar Tests for Each Category
 ------------------------------------------------------------------------
                                 Proportion
             Frequency           (Base Rate)
 Level    ----------------     ----------------       Chi-
  (k)     Path A     Path B     Path A     Path B     squared(a)     p
 ------------------------------------------------------------------------
   1         26         27       0.220     0.229     exact test    1.0000
   2         26         12       0.220     0.102        8.167      0.0043*
   3         38         69       0.322     0.585       27.457      0.0000*
   4         28         10       0.237     0.085       18.000      0.0000*
 ------------------------------------------------------------------------
 (a) or exact test
 * p < Bonferroni-adjusted significance criterion of 0.017.
                 Tests of Overall Marginal Homogeneity
     ------------------------------------------------------------
     Bhapkar chi-squared         =   38.528   df =  3   p = 0.0000
     Stuart-Maxwell chi-squared =   29.045   df =  3   p = 0.0000
                          Bowker Symmetry Test
            ------------------------------------------
            Chi-squared =   30.286   df =   6   p = 0.0000
```

Output

```
***TESTS FOR ORDERED-CATEGORY DATA***
                    McNemar Test of Overall Bias
                       or Direction of Change
             ----------------------------------------
             Cases where Path A  level is higher:      25
             Cases where Path B  level is higher:      18
             Chi-squared =    1.140   df = 1   p = 0.2858
 Four-fold tables tested (for thresholds tests)
    22     4     5    87
    36    16     3    63
    90     0    18    10
                     Tests of Equal Category Thresholds
-------------------------------------------------------------------------
           Proportion
            of cases
             below
            level k            Threshold(a)
Level   ----------------     ----------------       Chi-
 (k)    Path A    Path B     Path A    Path B     squared(b)     p
-------------------------------------------------------------------------
  2      0.220     0.229      -0.771   -0.743     exact test   1.0000
  3      0.441     0.331      -0.149   -0.439        8.895     0.0029*
  4      0.763     0.915       0.715    1.374       18.000     0.0000*
-------------------------------------------------------------------------
(a) for probit model
(b) or exact test
* p < Bonferroni-adjusted significance criterion of 0.017.
```

Output


GRAPHIC OUTPUT
Marginal Distributions of Categories
for Path A (**) and Path B (==)
0.585
0
1
2
3
4
Notes: x-axis is category number or level.
y-axis is proportion of cases.
Proportion of cases below each level
Path A
Path B
Scale
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.

Comments

- The Bhapkar test (p. 424, 11.3.1; more powerful than Stuart-Maxwell) for marginal homogeneity is highly significant with $p = 0.0000$. We reject marginal homogeneity. The graphical output indicates that both pathologists tend to classify 'negative' in roughly the same proportion, but that B classifies 'carcinoma *in situ*' more often than A, whereas A classifies classifies 'atypical squamous hyperplasia' and 'squamous or invasive carcinoma' more often than B.
- There is also an individual test for each category. $H_0 : P(X = i) = P(Y = i)$ is rejected for $i = 2, 3, 4$ but not $i = 1$.

Comments

- We are interested in whether one rater tends to classify slides 'higher' or 'lower' than the other. Off-diagonal elements above the diagonal are when B classifies higher than A; elements below the diagonal are when B classifies lower than A. The McNemar test of overall bias is not significant, indicating that one rater does not tend to rate higher or lower than the other.
- The test for symmetry has the same test statistic and p-value as from SAS.
- The program is easy to run on a Windows-based PC and free. There is a users guide and sample input and output files. Web location:

  `http://www.john-uebersax.com/stat/mh.htm`.

Stuart-Maxwell test in R

```
> library(coin)
Loading required package: survival
Loading required package: splines
> rate=c("N","ASH","CIS","SIC")
> ratings=as.table(matrix(c(22,5,0,0,2,7,2,1,2,14,36,17,0,0,0,10),nrow=4,
+  dimnames=list(PathA=rate,PathB=rate)))
> ratings
     PathB
PathA  N ASH CIS SIC
  N   22   2   2   0
  ASH  5   7  14   0
  CIS  0   2  36   0
  SIC  0   1  17  10
> mh_test(ratings)
        Asymptotic Marginal-Homogeneity Test
data:  response by
         groups (PathA, PathB)
         stratified by block
chi-squared = 29.0447, df = 3, p-value = 2.192e-06
```